CANELC: constructing an e-language corpus

Dawn Knight,[1] Svenja Adolphs[2] and Ronald Carter[2]

Abstract

This paper reports on the construction of the Cambridge and Nottingham e-language Corpus (CANELC).[3] CANELC is a one-million word corpus of digital communication in English, taken from online discussion boards, blogs, tweets, e-mails and Short Message Services (SMS). The paper outlines the approaches used when planning the corpus: obtaining consent, collecting the data and compiling the corpus database.

This is followed by a detailed analysis of some of the patterns of language used in the corpus. The analysis includes a discussion of the key words and phrases used, as well as the common themes and semantic associations connected with the data. These discussions form the basis of an investigation into how e-language operates in ways that are both similar to and different from spoken and written records of communication (as evidenced by the British National Corpus, BNC).

Keywords: blogs, tweets, SMS, discussion boards, e-language, corpus linguistics

[1] School of Education, Communication and Language Sciences, Newcastle University, Newcastle, NE1 7RU, United Kingdom.
[2] School of English Studies, University Park, University of Nottingham, Nottingham, NG7 2RD, United Kingdom.
Correspondence to: Dawn Knight, *e-mail*: Dawn.Knight@ncl.ac.uk

[3] This corpus has been built as part of a collaborative project between the University of Nottingham and Cambridge University Press with whom sole copyright of the annotated corpus resides. CANELC comprises one-million words of digital English taken from SMS messages, blogs, Tweets, discussion board content and private/business e-mails. Plans to extend the corpus are under discussion. The legal dimension to corpus 'ownership' of some forms of unannotated data is a complex one and is under constant review. At present, the annotated corpus is only available to authors and researchers working for CUP and is not more generally available.

Corpora 2014 Vol. 9 (1): 29–56
DOI: 10.3366/cor.2014.0050

www.euppublishing.com/cor

1. Introduction

Communication in the digital age is a complex, many-faceted phenomenon involving the production and reception of linguistic stimuli across a multitude of platforms and media types (see Boyd and Heer, 2006: 1). While a wealth of corpus research has been carried out on individual forms of e-language (i.e., language communicated through any digital device), from SMS messages, to blogs and e-mails, the corpora that have been utilised to date tend to be either small-scale or bespoke, that is, planned or utilised to answer very specific linguistic enquiries, and/or they consist of only one e-language variety (see Puschmann, 2009; Schler *et al*., 2006; and Tagg, 2009; for examples).

The most notable, large-scale selection of examples of current e-language corpora are detailed in Table 1. While invaluable for examining language patterns of their own individual text-type/language variety, such corpora are limited in their utility. Although it is widely acknowledged that we live and communicate in a digital world, the ways in which we do this, across multiple resources, remains an under-explored area of research in corpus linguistics: there is a lack of appropriate existing resources to make this work possible.

The next phase of corpora development should, therefore, seek to fill this void and integrate a wider range of different and relevant digital resources for linguistic analysis in a large-scale functional database. Arguably, a seemingly logical way to develop an integrated e-language corpus would be to attempt to combine all current corpora and, thus, build a corpus of existing legacy data. While this would, in theory, allow us to construct a large-scale corpus in very little time, (with, it is assumed, minimal effort), in reality, various practical and ethical challenges would be encountered in the process.

First, it is likely that each of the corpora have been constructed in a different way, using different methods for extracting and storing data, with different header and related information being retained in each of them. Furthermore, the Blog Authorship corpus, for example, is already a few years old, so the content is relatively outdated. This may limit the extent to which we can use analyses from this corpus to discuss patterns of e-language use in the 2010s, and it also limits the comparability to the other e-language corpora in existence, since they all contain data from different time periods.

In order to ensure that the content of the corpus is structurally and compositionally consistent, accurate and as up-to-date as possible, it was deemed more viable to construct a new e-language corpus from the bottom up – one that, preferably, includes data from the past one to two years. The remainder of this paper introduces one such corpus, the newly constructed Cambridge and Nottingham e-Language Corpus (CANELC) corpus. We outline the basic composition of the corpus, the approaches used in recruiting contributors and compiling the data, and then provide some results from preliminary analyses of the data, focussing specifically on defining some

Name	Description	Reference
Blog Authorship Corpus	Freely available 140 million word English blog corpus. Comprises 681,288 blog entries taken from 19,320 bloggers over three different age groups	Schler *et al.*, 2006
CorText	110,000 word corpus of SMS messages (19,000 messages from 235 users), also with associated biographical metadata	Tagg, 2009
Dortmund Chat-Korpus	One-million tokens from online chatrooms in German (140,000 chat conversations)	Beißwenger, 2007
Enron Corpus	Seventy-million words from e-mails sent by 150 individuals, mainly senior managers of the ENRON firm	Klimt and Yang, 2004
Junk Email Corpus	373,000 words from 1563 junk e-mail messages	Orasan and Krishnamurthy, 2002
NPS Chat Corpus	10,567 messages from online chatrooms in English	Forsyth and Martell, 2007
Twitter_Smallcorp	Two-million word corpus of tweets	Puschmann, 2009
ICWSM Conference (TREC Tweets 2011)	Sixteen-million word corpus of tweets	Horn *et al.*, 2011

Table 1: A selection of current e-language corpora

patterns in function, sense and meaning used in the corpus and how these compare with spoken and written components of the BNC.

2. Why build an e-language corpus?

The motivation for building CANELC was twofold. Firstly, from the perspective of Cambridge University Press, it was conceived as a potentially

invaluable teaching and learning resource. It has been designed, therefore, with the purpose of informing and supporting content included in text books and grammars published by the Press, with specific extracts of the corpus used to illustrate ideas and/or specific functions and properties of language usage. Secondly, our own academic interest in CANELC encompasses a broader range of concerns: it not only provides the facilities for exploring patterns of lexical, grammatical and semantic properties of language use within and across different communicative modes, but it also helps us to develop our understanding of how these patterns of usage compare to and contrast with those seen in previous corpus-based studies of spoken and/or written discourse. We are aware that a tweet or a thread on a discussion board, for example, is lexically and structurally different from standard written and spoken English; but exactly *how* and *why* they are different (and in which ways: from each different text type to the next) are questions that have yet to be explored fully.

Analyses of the Cambridge and Nottingham Corpus of Discourse in English (CANCODE)[4] spoken corpus, in comparison with written counterparts from the Cambridge English Corpus[5] (CEC), for example, indicated that spoken language exhibits a marked increase in the use of personal pronouns, discourse markers and response tokens in comparison with written language (see Carter and McCarthy, 2006: 9–16). So words such as *it's*, *yeah*, *I*, *you*, *you know* and *mmm* are indicative of spoken discourse and are considered to be markers of interactive informality. More formal linguistic structures such as *whom* and *no one* (in contrast with *nobody*, as is often used in spoken language), were found to be comparatively more frequent in the written components of CEC.

Crystal (2003: 17) suggests that spoken and written language exist on a continuum of formality. The more formal language structures and conventions exist at the written end of the spectrum and the least formal exists towards the spoken end. The question is where forms of e-language exist on this continuum. Crystal suggests it is perhaps somewhere in the middle, as a distinct form of language in itself, but where exactly this lies is still under debate by researchers in the field. It is to be hoped that analyses of data from CANELC will allow us to make better-informed judgments on the nature and characteristics of e-language and its 'best fit' along the continuum, providing the foundations for better-enhancing our descriptions and understanding of language use in the modern digital age.

The issue of levels of formality in specific types of e-language has already received attention from researchers (see works by Crystal, 2008; Hård af Segerstad, 2002; Shortis, 2007; and Sutherland, 2002, for further details). Tagg (2009) and Ling (2003) both report on the tendency for most

[4] CANCODE contains five-million words of (mainly casual) conversation recorded in different contexts across the British Isles.

[5] The CEC contains over one-billion written and spoken words in English. For more information, see: http://www.cambridge.org/

SMS messages to be immediate and personal, written in the first person and directed to specific recipients. Tagg adds to this, suggesting that 'the informal and intimate nature of texting encourages the use of speech-like language' (2009: 17; also see Oksman and Turtianen, 2004). Similarly, Baron (1998: 36) argues that e-mail, as with texting and other common forms of e-language, is a written mode of communication but that 'participants exploit it for typically spoken purposes'; it perhaps shares, therefore, more similarities with communication situated at the spoken rather than the written end of the continuum. The blurring of traditional characteristics of spoken and written language in digital communication is something that has been discussed at length, although there is a limit to which this has been supported by corpus-based analysis of real-life data (see Biber, 1993; Collot and Belmore, 1996; and Crystal, 2001, for further details). The CANELC corpus enables such investigation.

The level of formality in text is closely related to the function of the message, and this poses a variety of challenges when classifying text types since non-typical features may be included – perhaps as an expression of creativity or style or because the 'context' in which they are used causes these changes in language use (for discussions of language and context, see Bates, 1976; Brown, 1989; Duranti and Goodwin, 1992; Green, 2002; Halliday and Hasan, 1989; Labov, 1972; Nelson *et al*., 1985; Scollon and Scollon, 2003; and Widdowson, 1998). Who sends a message to whom and when and where this occurs can impact on the meaning and the pragmatic function of the content. SMS messages sent from a business director to a managing director of a different company, for example, are likely to be more formal than a message between two friends arranging a coffee date. Given this tendency, information such as the date and time messages were sent, and the identity of senders and recipients (including age, gender, occupation, nationality and relationship), are relevant and should be retained as metadata when constructing new corpora of this nature. This information can then be consulted when analysing the data in order to help to reconstruct elements of the fragmented context of the language in use and to help to explain why certain patterns may exist.

It is unlikely that complete metadata records of e-language contributors can ever be constructed, because users often adopt a certain level of anonymity when online (especially when sharing data) and engaging in forms of e-communication. Furthermore, the notion of 'context' is in itself difficult to define and qualify, as is the extent to which it shapes or develops the meaning of a message. This is because context is a complex, fluid notion that involves social, physical and temporal dimensions which are often abstract. For example, a location may be defined by the use of the absolute – a specific grid reference on a map, street X. To an individual who is standing in street X, perhaps sending an e-mail or writing a text, street X may be the location of a public house or a pool hall, a place where the contributor visits with specific friends or colleagues, meeting at certain times of the week to partake in a particular activity, for example. Understanding and accounting for these

more complex structures of the social context of the message will allow us to enhance our descriptions of language in use, providing an insight into more pragmatic, functional aspects of communication.

In practice, however, it is unlikely that such enriched information can be gathered successfully when constructing large-scale corpora, as such detailed enriched profiling of language is only really practical on a small-scale with the involvement of limited numbers of contributors. Despite this, it is still important that we at least aim to collect some basic forms of metadata, including biographical information about contributors, where they are in the world, and categorising the intended readership of content: this information may still prove to be of interest when examining the corpora in more detail.

3. Composition of the corpus

3.1 Data included

There are a range of different e-language resources that are used as a means of communicating in everyday life, from SMS messages to e-mail activity, blogs, status updates on social networking sites and instant messaging conversations. CANELC contains the data shown in Table 2.[6]

As outlined by Herring (2002), there is a variety of ways of classifying computer-mediated discourse. For the purpose of CANELC, the data is broadly categorised under a range of different 'genres' (Herring, 2002) with the overarching grouping of 'e-language'. These genres are essentially individual 'socio-technical modes', each of which is likely to have specific 'social and cultural practices that have arisen around their use' (Herring, 2007: 3). Coupled with the addition of metadata detailing not only the specific mode of language, but also information of the 'participant characteristics', 'topic or theme' and so on (Herring, 2002: 19, see Sections 5 and 7), this broad method of categorisation provides a way-in to exploring patterns of language use and to carrying out corpus-based linguistic research at the communicative *mode* level.

CANELC was also constructed to allow for querying the data, at a more general level, of the genre of communication (that is 'e'-based language or 'netspeak'; see Crystal, 2001). This is because, despite being *different*

[6] Externally commissioned research is, to some degree, always subject to the requirements of the agency that commissions the research, and the balance of CANELC data is determined accordingly with SMS and e-mail datatypes assuming a smaller proportion. The next phases of the research may, indeed, see each of the data-type categories balanced more evenly. However, SMS and e-mail data are categorised by a markedly interpersonal dimension and when aggregated do constitute a further balancing category in the whole corpus.

Data type	Number of contributors	Number of messages/ entries	Word count	
			Raw	Percent
Twitter	30	18,972	259,101	26
Blogs	36	1,101	267,983	27
Discussion boards	12	2,715	242,727	24
E-mails	various	1,920	128,951	13
SMS	11	5,215	101,913	10
		29,923	1,000,675	100

Table 2: The contents of CANELC

socio-technical modes, there is a key similarity between them to the extent that they are all forms of asynchronous communication systems. Although SMS and e-mail constitute interpersonal communication exchanged between a potentially large but bounded number of participants (discussed further in Section 4), and Twitter and blogs are, instead, usually publicly accessible, none of these different modes 'require that users be logged on at the same time in order to send and receive messages' (Herring, 2007: 13). Instead, as with most written forms of language, these 'messages are stored at the addressee's site until they can be read' (Herring, 2007: 13). This is different to spoken language which is, conversely, most often synchronous.

Herring underlines that this makes 'synchronicity a useful dimension for comparing different types of CMC [computer-mediated communication] with spoken and written discourse' (Herring, 2007: 9; see also Condon and Cech, 1996; and Ko, 1996) and, although specific differences in patterns across the individual 'modes' are underlined, it is this dimension that motivates the preliminary comparisons between spoken, written and 'e-language' that are carried out in the final part of this paper.

3.2 Recruitment and permissions

To collect data for CANELC, authors of 'popular' blogs, discussion boards and tweets were targeted as it was thought that this would best represent the types of the discourse that the general public would be reading. This notion of 'popularity' was gauged according to the following requisites:

- Sites had to feature within online directories[7] of the most popular blog/tweet lists (sourced by Googling 'top ten blogs', 'popular blogs' and so on);
- Tweeters were to have at least 1,000 followers; and,
- Posts from blogging sites had either a range of readers/followers and/or numerous responses to posts, indicating a large readership.

These were sites with 'public' rather than private profiles. From these lists, the specific individuals contacted (as hundreds of individuals were included here) were chosen at random in the first instance and were then filtered further by means of checking whether the following criteria were met:

- The prospective site was managed by a single individual (to ease problems associated with permissions for multiple contributors), who assumed copyright for their own material;
- E-mail/contact details were easily obtainable; and,
- The site contained a reasonable amount of text.

3.3 Gaining permission

Hundreds of potential target sites were shortlisted using this approach. Contact details of the owners/moderators of the sites were tabulated, with each being contacted to ask for permission to use their data. The permission process was tested during the piloting phase. The initial approach was to send a traditional consent form attached to an e-mail detailing the aims and objectives of the study, then requesting each individual to provide permission in a response to the e-mail, then to sign the form and return it to the researcher.

Thirty prospective blog and Twitter contributors were contacted during this piloting phase and while twelve individuals responded, only seven of these provided full permissions. Five others declined to participate and the remainder neglected to respond. The positive response from the twelve individuals was reassuring but it was decided that a more streamlined approach for providing consent was needed as the process of posting and/or

[7] Examples of such sites include: www.guardian.co.uk/technology/2008/mar/09/blogs
http://modernl.com/article/uk-blogosphere-top-10-british-blogs
http://wefollow.com/twitter/british
www.britishblogs.co.uk
www.telegraph.co.uk/technology/twitter/6832287/Most-influential-British-twitter-users-revealed.html

scanning a long and detailed form was time-consuming and inconvenient. As a second parse, instructions regarding the provision of consent were written into the initial correspondence sent to prospective contributors, thereby streamlining the process. This allowed individuals to simply respond with 'yes, I provide consent', without having to go through the more laborious form-signing process.

Striving for consistency in the type of correspondence and documentation sent to each prospective contributor was of paramount importance to this project given that extracts of the corpus are likely to be published in academic texts and teaching materials. Therefore, in consultation with CUP and their legal team, an e-mail and permission form was drawn up in order to verify the legitimacy of the permission sought. This was circulated to over one-hundred potential sites/individuals and in instances where 'full' permission was granted, data was sampled. Permissions were not sought, and data was not taken from third parties who, for example, responded to content on a blog. However, a note of how many responses were associated with specific contributions *was* made in order to enrich the dataset.

With the discussion board data, consent was requested in the same way, but, as an additional measure, discussion board moderators were asked with whom the sole copyright of content lay. If it was with the moderators themselves, content was taken from all users who had made a contribution to the board. If not, individual contributors as well as the moderator were contacted and asked for permission to use their text. Again, only text provided by fully consenting moderators and/or individuals was used in CANELC.

3.4 Profile of contributors

CANELC aimed to include contributions from a range of different sociolinguistically profiled participants (that is, of different ages, genders and so on). As far as possible, requisites identified in the 'aimed composition' column of Table 3 were to be met so as to ensure balance and consistency in the data. The 'actual composition' column of this table describes the extent to which these were met.

3.5 Access

Initial plans were to make this corpus open access. Unfortunately, ownership and distribution rights, enforced by our partners, have resulted in restricted access to the corpus. It is, thus, only available to researchers at the University of Nottingham or staff working at CUP.

Variable	Aimed composition	Actual composition
Number of participants	10–40 per source	11–36 contributors per source
Gender	50:50 male and female	50 percent of the corpus has a circa 50:50 balance. For 50 percent genders are unknown
Age	Under 19, 10 percent of total 20–24, 10 percent of total 25–29, 10 percent of total 30–34, 10 percent of total 35–39, 10 percent of total 40–44, 10 percent of total 45–49, 10 percent of total 50–54, 10 percent of total 55–59, 10 percent of total Over 60, 10 percent of total	Contributors were from a range of different age groups although the most populous groupings were 20–24 and 25–29 (there was not a strict balance of contributions across the groupings)
Time period	Contributions posted from 2006–2011	Data from each contributors was collected over a minimum of three days, the majority within the 2010–2011 period
Location	100 percent posting to sites ending in .co.uk	All sites ended in .co.uk and most contributors identified themselves as being British

Table 3: Profile of the contributors to CANELC

4. Data types[8]

4.1 Tweets

It is estimated that over 175 million people use Twitter[9] globally, to update their 'followers', friends, and/or the world at large on their thoughts, feelings

[8] Facebook status updates are not included in CANELC as the fact they can be viewed and commented on by an array of different users ('friends'), commonly commenting on private information about the user and his/her friendship group, brought into question concerns about copyright and data ownership. To avoid problems with access, ethics and copyright, this data was not included.

[9] See: www.twitter.com

and reflections at a given moment. It is often used in a professional capacity, for publicity or advertising, but is also used on a more personal level, for individual tweeters to talk about their daily lives. Twitter operates in a similar way to Facebook[10] updates and SMS messages in that it is restricted in terms of the number of characters (140) that can be inputted on a tweet at any one time. But a 'tweeter' has no restrictions on the number of messages they can post over the course of a day.

An increasing number of linguistic studies have been carried out on the language of tweets (for example, see Borau *et al*., 2009; Honeycutt and Herring, 2009; Jansen *et al*., 2009; and Zappavingna, 2011) and, as identified in the introduction, there is an increase in interest in building and using Twitter corpora, particularly in the field of Natural Language Processing (NLP), for the purpose of sentiment analysis.[11]

Tweets, in the same way as our second data type, blogs, can be classified as 'outward facing' forms of digitally based communication: they are posted on sites which can be accessed by anyone (unless they are hosted on member only sites) and so, it can be assumed, are aimed at a wider readership and audience than a personal SMS or Instant Message (IM – another form of e-language). The readership is often less specific, although the content of the material may be of interest to some individuals more than others. For example, a middle-aged university lecturer may be more interested in the content posted by a publishing house, research network or fellow academics, rather than that posted by Britney Spears or the pop group, JLS.

We faced a challenge when trying to decide which tweeters to target when constructing the CANELC corpus. We wished to collect data which was as 'representative' of each different e-language type as possible, rather than simply using a web-crawler or API to collect data, randomly selecting sources. To achieve this, we decided to collect data from popular public sites only (see Section 3.2) – ones that discuss a range of different topics, have as large a readership as possible and whose authors provided full permission to reuse their data. The selection and classification of topics was consistent to the approach used when collecting blog and discussion board data, as defined in Section 5.

4.2 Blogs

Twitter was only launched in 2006 so the use of tweets has a relatively short history; but the use of weblogs (blogs) saw a 'sudden rise in prominence in 1999' (Myers, G., 2010: 10) and they are now authored by billions of web users across the globe. Blogs are generally longer excerpts of prose

[10] See: www.facebook.com

[11] For examples, see: http://www.tweetfeel.com/, www.sentiment140.com, http://tweetsentiments.com and http://www.tweettone.com/

as they are not restricted by space or word count, so can run from a few sentences to numerous paragraphs of text. 'Blogging software means that anyone with access to the internet can post their thoughts, links and photos on a blog' (Myers, G., 2010: 77) although the readership of a given blog is again dependent on who is writing it, the topics covered by the content, accessibility to the content and how the blogs are presented.

There is an ever-increasing amount of research being carried out on blogs. One key area of study has focussed on exploring patterns of language use and the social functions of blogging (see Allan, 2006; Gillmor, 2004; Myers, B., 2010; and Myers, G., 2010). The inclusion of blogs in CANELC aims to complement this existing research. It also aims to allow us to examine the relationship between this mode of e-language and other varieties.

4.3 Discussion boards

Discussion boards are more interactive spaces for online communication. In a similar way to IMs and interaction on social networking sites (SNSs), individuals add comments about a given topic, either prescribed by the site moderator or by the first contributor to a thread, and others read and respond to the comment through supporting, challenging and/or building on what has been said. Research on the social dynamics of Internet forums has been widely published, often exploring the notion of the generation of a 'virtual community' through language (Jones, 1997) in a 'virtual space' (Rheingold, 1993). An example of this includes a recent thesis by Atkins (2011), exploring the indexing of space through the use of language (mainly deixis) in Internet health groups using a 45,000 word corpus. Before CANELC, a corpus including threads from a wide range of discussion boards, covering a broad spectrum of different topics, had yet to be compiled.

4.4 E-mails

E-mails are often only addressed to specified recipients or groups of readers, and are not outward facing or designed for the public at large, although the number of potential recipients of an e-mail may actually be infinite. In a similar fashion to IM content, users can respond in a chain-like fashion to previous messages, with as little or as much text as they choose and whenever and wherever they like, through a PC/laptop or mobile phone.

Research into the language of e-mails is again longstanding; noteworthy examples include works by Baron (1998, 2000), Crystal (2001), Danet (2002) and Panteli (2002). As with the other e-language types, e-mail corpora are constantly emerging, and the content of the large-scale Enron corpus, most notably, has been studied in some detail by researchers in this field already. Despite such work, the similarities and differences between

e-mails and other forms of e-language, in terms of structural, functional and pragmatic properties, remains underexplored. CANELC gives researchers an impetus for carrying out such lines of research, as well as for building on the foundations of what is already known about the language of e-mail. It should be noted, however, that our data consists largely of e-mail collected from business contexts. It is not especially representative. The request from CUP, however, was for us to collect business data to inform text and course book development in business English. A next stage would be to collect a greater variety of less contextually specialised e-mails.

4.5 Text messages

The final form of e-language included in CANELC is SMS messages. While 'text messaging was never originally envisioned as a means of communication between individuals … it was originally conceived of as having commercial use, or possible as a service for mobile phones to signal the arrival of a voicemail message' (Crystal, 2008: 77), 'texting' has become a very central part of communication in modern life, with eleven-million text messages being sent every hour in the UK (as recorded in January, 2010).[12]

SMS messages are, again, more private forms of communication as they are often directed at individuals and small groups of friends. Texting is immediate and often informal. The language of SMS messages has been explored by numerous researchers in linguistics (see Crystal, 1998; Döring, 2002; Faulkner and Culwin, 2005; Grinter and Eldridge, 2003; Tagg, 2009; and Thurlow and Brown, 2003) although, as Crystal (2008: 28) notes, 'we are still learning how to behave when we text'. Issues such as when and how one should appropriately respond to messages, if at all, are widely debated. There is, thus, scope for examining other characteristics of SMS behaviour that are still somewhat underexplored, and, again, the provision for doing just this is something that CANELC offers.

5. Topics covered

CANELC includes data covering a range of different topics. A list of these is provided in Figure 1. We originally intended to use pre-existing schema to describe and encode the different topics of the content, but we were unable to find a generic, widely used classification system for this purpose. Therefore, these categories were defined following extensive discussions by the group of researchers working on the CANELC project. The team looked at the key content words in the descriptions of sites, such as 'food' and 'recipes' in 'Showing the world the beauty of British food and recipes',

[12] Figure taken from findings of The Mobile Data Association: http://www.themda.org/ (accessed 1 June 2011).

Figure 1: Topics covered in the CANELC content

noted them down for each individual contributor, then attempted to define broad thematic categories based on the key words defined across the dataset. So, for the example just given, text collected from this tweeter would be broadly categorised under the topic of 'cookery', so would be labelled under category 'C'.

The categorisation of the data was carried out semi-automatically. As a first parse, two trained researchers were employed to look at the data and categorise topics manually. The data were then inputted into the semantic tagger of Wmatrix (Rayson, 2003) to see whether thematic groupings of the content could be defined using this automated method. Finally, the results from the three stages were compared and lengthy discussions were carried out between the researchers to select which category appeared to be the most relevant to specific extracts of data. In situations where differences of opinions could not be resolved, as discussed above, multiple codes were assigned to the data rather than one.

These 'topics' exist on a continuum from the more 'public' concerns (topics in the 'A' category) such as news, politics and current affairs, to the 'private', such as personal and daily life. The entire CANELC dataset has been categorised according to these categories. While the assignment of the content to these categories was fairly transparent in some cases, others were slightly more 'fuzzy' in that they discussed multiple topics across the different categories. In these instances, the data were given a range of category codes, thus A/B/C rather than simply 'A'.

6. Anonymity

To protect the identity of contributors to the corpus and individuals mentioned within it, all content has been fully anonymised. First names (including Twitter IDs, *etc*.) and easily identifiable nicknames were anonymised as [NameX], with 'X' representing a unique number code which is indexed in our metadata files (though these are unlikely to be distributed). Other anonymised features/codes include the following:

[Address]	[Link]	[ServerAddress]
(ContentPrivate)	[Password]	[ServerName]
[Bankdetails]	[PhoneNo]	[Signature]
[BusinessName]	[Photo/Picture]	[SoftwareName]
[DocumentRef]	[PONumber]	[Sortcode]
[Email]	[PortNo]	[Tagline]
[FaxNo]	[ProductName]	[Username]
[IPAddress]	[Postcode]	[Website]

Anonymising e-language is a challenging process. This is especially true for the shorter and fragmented contributions such as SMS messages and tweets. This is because references to persons/ places in such discourse tend to be highly context-bound and are, thus, integral to the meaning of the message, making it potentially detrimental to remove them. For this reason, the same approach used by Tagg when developing the CorText corpus was used here, wherein text referring to 'celebrities, film names, public places, characters from film, TV/ reality shows weren't changed' (Tagg, 2009: 98) but references to persons who are not in the public eye, along with those other features mentioned above, were.

Given that 'popular' blog, discussion board and Twitter sites were included in the corpus, such public figures, for example, featured frequently. An example of this is seen in the following tweet:

Sent – 22:17 on 12/12/2010
Content – @[Name1911] Feel exactly the same. Old school Biffy fan, Essex born and Matt fan but I'm conflicted

'Biffy' in this tweet refers to the band Biffy Clyro, whose song was covered by 'Matt', a singer from Essex who won the TV show X Factor in December, 2010. Without this extra information – that is to say, the name and identity of the band/singer – the analyst would be unaware of the referential meaning of this tweet. For this reason, details of this nature, such as public figures, designer labels, celebrities, TV programmes/characters and shop names were not anonymised.

To add clarity and extra meaning to such contextually bound referents, an index of unanonymised content was created while constructing the corpus, detailing the name of the referent and who/what they are. An excerpt from this 'index of cultural referents' is seen in Figure 2.

Additional Details (not anonymised):		
Twitter	ABBA	Band
Twitter	ABC	American TV network
Twitter	Alastair Cook	England Cricket Player
Discussion Boards	All-American Rejects	Band
SMS	Alton towers	Leisure Park
Discussion Boards	Android	Mobile phone variety
Twitter	Andy Murray	Tennis Player
Twitter	Anfield	Liverpool FC home pitch
Twitter	Angry Birds	Electronic game
Discussion Boards	Ars Technica	Technology brand
Twitter	Avatar	Film
Discussion Boards	Axel Rose	Lead singer of rock band Guns N Roses

Figure 2: An example of the index of unanonymised content in CANELC

Common Christian names and nicknames were also, in some cases, not anonymised because it was felt the identity of the referent could not be easily traced when using such names. An example of this is seen in the following tweet:

Tweet T.12115 @[Name2856] Thx 4 the RT Andy.

It is unlikely that the identity of the specific person 'Andy' can be determined simply by reading this tweet, so it was felt that it would not be cause for concern to leave such names in the data.

7. Storing and representing the data

When permission was granted, data were simply extracted from the site(s) or RSS feeds (for blogs, discussion boards and tweets) and pasted into an XML corpus database. This database was standardised and formatted in the same way as content from the Cambridge English Corpus so that the data can be directly slotted into this corpus.

As far as possible, data were stored with the following headers included:

- Author's name, age, gender and nationality
- Date and time composed
- Intended recipient
- Content
- General theme of content
- Follow up comments/responses
- 'Other' relevant information

Data from e-mails was forwarded directly to the researcher who could input the content into the XML database manually. Many modern smart

phones are compatible with PC based software which allows users to connect their phones with USBs to computers and simply download the content of their SMS messages along with the time and date they were sent. Alternatively, web-enabled phones often automatically back-up these details to web accounts which can be downloaded as a database and sent directly to the researcher for use.

8. Analysis

8.1 Key questions

The final part of this paper reports on some preliminary analyses of CANELC. The aim, here, is to outline some of the basic similarities and differences between the asynchronous data included in CANELC and one-million word samples of spoken and written language from the BNC.[13] The following key question is examined as part of this analysis:

> What are the most frequent words/phrases used in CANELC (within and the across different modes) in terms of word *function*, *sense* and *meaning* and how do these compare to the spoken/written elements of the BNC?

For the purposes of this analysis, non-standard spellings featured in the corpus, such as *2* (for 'to' or 'too'), *wanna* ('want to') and *u* ('you') were standardised with the help of the software known as VARD (Baron and Rayson, 2008), prior to being inputted into Wmatrix. The corpus data was grammatically and semantically tagged in Wmatrix after the standardisation had taken place. VARD enables users to identify spelling irregularities in a corpus then train the system to replace candidates with standardised versions of the words automatically. These were then counted towards cases of the standard spelling of each form. Given that the orthographic formulation of these features was not the primary concern of the analysis, (rather, the frequency of use with which forms were used), this process of standardisation was deemed sufficient for the needs of this paper.

8.2 Function

Figure 3 tabulates the top fifty most frequent words and clusters used in the CANELC corpus (note: 'rel. freq.' refers to relative frequency; this is the frequency of a given word as a proportion of the entire corpus).

---

[13] The BNC is a 100 million word corpus of written and spoken discourse in English. For more information, see: http://www.natcorp.ox.ac.uk/

No.	Word	Freq.	Rel. freq.	No.	Word	Freq.	Rel. freq.	No.	Word	Freq.	Rel. freq.	No.	Word	Freq.	Rel. freq.
1	the	14391	3.96	26	if	1655	0.46	1	going_to	291	0.08	26	last_night	65	0.02
2	I	8446	2.33	27	will	1565	0.43	2	thank_you	286	0.08	27	next_week	61	0.02
3	to	8424	2.32	28	at	1528	0.42	3	have_to	247	0.07	28	set_up	54	0.01
4	a	7810	2.15	29	so	1511	0.42	4	let_know	178	0.05	29	do_you_think	54	0.01
5	and	7294	2.01	30	me	1464	0.4	5	a_lot	149	0.04	30	looks_like	52	0.01
6	of	5674	1.56	31	as	1424	0.39	6	a_bit	149	0.04	31	make_sure	51	0.01
7	you	5237	1.44	32	from	1399	0.39	7	as_well	135	0.04	32	your_own	49	0.01
8	@	4636	1.28	33	can	1342	0.37	8	out_of	117	0.03	33	for_example	48	0.01
9	it	4549	1.25	34	just	1300	0.36	9	at_least	110	0.03	34	last_year	48	0.01
10	in	4217	1.16	35	your	1285	0.35	10	credit_card	105	0.03	35	at_the_moment	46	0.01
11	is	3952	1.09	36	all	1209	0.33	11	a_couple	102	0.03	36	looking_at	46	0.01
12	for	3763	1.04	37	they	1178	0.32	12	of_course	102	0.03	37	up_to	45	0.01
13	that	3633	1	38	x	1126	0.31	13	a_little	98	0.03	38	very_much	43	0.01
14	on	2874	0.79	39	or	1125	0.31	14	check_out	96	0.03	39	so_far	40	0.01
15	be	2543	0.7	40	would	1053	0.29	15	look_at	93	0.03	40	due_to	40	0.01
16	have	2410	0.66	41	what	1052	0.29	16	rather_than	92	0.03	41	out_there	40	0.01
17	with	2265	0.62	42	by	1007	0.28	17	i_think	89	0.02	42	even_if	40	0.01
18	this	2222	0.61	43	an	997	0.27	18	had_to	81	0.02	43	at_all	40	0.01
19	are	1995	0.55	44	there	992	0.27	19	make_it	77	0.02	44	such_a	40	0.01
20	but	1954	0.54	45	one	984	0.27	20	used_to	75	0.02	45	sounds_like	39	0.01
21	do	1863	0.51	46	about	938	0.26	21	you_know	75	0.02	46	too_much	39	0.01
22	my	1848	0.51	47	get	916	0.25	22	new_year	73	0.02	47	such_as	39	0.01
23	not	1769	0.49	48	some	874	0.24	23	looking_for	69	0.02	48	down_to	38	0.01
24	was	1762	0.49	49	more	844	0.23	24	this_morning	66	0.02	49	in_order_to	37	0.01
25	we	1744	0.48	50	has	840	0.23	25	looking_forward	65	0.02	50	so_that	36	0.01

Figure 3: The most frequent words and clusters used in the CANELC corpus

Function words, rather than content words, proliferate here. Of these function words, we find that, significantly, the use of personal pronouns is shown to be particularly frequent in the data, with *you*, *I* and *it* ranking highly, along with the demonstratives *this* and *that*. A keyword analysis of these pronouns, in comparison to spoken and written extracts of the BNC (comprising one-million words each), indicated that their use more closely mirrors spoken forms of discourse, as personal pronouns are characteristically less often used in written language. *I*, for example, was noted to occur once every thirty-eight words in an analysis of some spoken data from the BNC (Leech, 2000) and only once in every 200 words in the written data. Rates of 1:43 words for the CANELC data are, thus, far closer to the spoken BNC analysis. (This result was also mirrored by Atkins, 2011; Biber, 1992; Biber *et al*., 1999; Carter and McCarthy, 2006; and Chafe and Danielewicz, 1987.)

As outlined by Heylighen and Dewaele (2003), the frequent use of personal pronouns, along with adverbs, verbs and interjections is typically characteristic of more informal styles of communication, while nouns, adjectives, prepositions and articles are more frequent in more formal types of language. Based on this basic definition, and to provide an insight into the levels of formality across the different communicative modes in the corpus (albeit crudely), Figure 4 shows the relative frequencies of each of these parts of speech across the modes in the corpus, compared to the relative frequencies seen in the entire corpus.

The numbers in bold indicate that the relative frequency for a specific part of speech is lower than that seen across the entire CANELC corpus, while those in italics are higher than the overall relative frequency for the corpus. Blog and Twitter data are shown to use parts of speech that are more

Part of speech	CANELC	Blogs	Tweets	Discussion boards	Email	SMS
Pronouns	9.03	**7.57**	**8.02**	*10.58*	*10.48*	*9.4*
Adverbs	6.85	**6.79**	**5.67**	*6.95*	*7.5*	*8.94*
Verbs	20.45	**18.69**	**17.97**	*22.29*	*23.23*	*23.65*
Interjections	0.59	**0.18**	*0.79*	**0.3**	*0.94*	*1.73*
TOTAL	36.92	**33.23**	**32.45**	*40.12*	*42.15*	*43.72*
Nouns	22.77	*25.23*	*25.3*	**21.01**	**19.47**	**18.49**
Adjectives	6.9	*8.34*	**6.7**	**6.47**	**5.74**	**6.09**
Prepositions	10.11	*10.56*	*12.97*	**8.71**	**8.5**	**7.28**
Articles	6.67	*8.51*	**4.8**	*7.08*	*6.72*	**4.37**
TOTAL	46.45	*52.64*	*49.77*	**43.27**	**40.43**	**36.23**

Figure 4: Relative frequencies of syntactic categories across the CANELC corpus

characteristic of 'formal' language at a higher rate than other modes across the corpus, while discussion boards, e-mails and SMS messages are closer to more 'informal' language. The most significant differences in relative frequency are seen in the underuse of nouns in the Twitter and blog data, the underuse of verbs in the e-mail data and the overuse of verbs in the Twitter data.

A variety of reasons may account for these differences, many of which are likely to be closely tied to 'social factors associated with the situation or context of communication' (Herring, 2002: 11; also see Baym, 1995; and Hymes, 1974). Content sent/received through the blogs and tweets (particularly those selected to be included in this study) is often publicly rather than personally targeted. This means that the readers are often unknown, so the relationship between the blogger/tweeter and reader is often less close than it is with SMS users. The 'purposes of communication' may also be different from e-mails and SMS messages which, again, in turn affects what they are communicating about and how they achieve this (i.e., the type of language being used). A closer analysis of these social factors and the individual context of communication will allow more specific conclusions about this to be made. Again, the detailed metadata associated with the CANELC data will allow us to explore this further in future studies; there is, however, limited scope to do this here.

SMS and e-mail are often more personal and intended for a bounded number of recipients. The language used in these situations may still be formal, such as in professional e-mails for example, but the recipient is often more likely to be a known party or somebody in close proximity to the sender's social or peer group.

8.3 Sense and meaning

Figure 5 lists the top fifty key words and clusters used in the CANELC corpus, compared to spoken and written BNC extracts.

CANELC Vs Spoken BNC								CANELC Vs Written BNC							
No.	Word		LL. Score	No.	Word		LL. Score	No.	Word		LL. Score	No.	Word		LL. Score
1	@	+	12146.53	26	let_know	+	330.59	1	@	+	12046.47	26	it	+	423.48
2	x	+	2260.52	27	love	+	314.19	2	i	+	5219.64	27	im	+	406.82
3	thanks	+	984.52	28	by	+	298.75	3	you	+	3242.12	28	love	+	373.07
4	u	+	933.01	29	great	+	290.74	4	x	+	2097.52	29	thank_you	+	363.15
5	hi	+	932.68	30	thats	+	288.2	5	u	+	1673.41	30	think	+	356.69
6	xx	+	686.45	31	from	+	285.23	6	thanks	+	1383.82	31	have	+	335.51
7	its	+	649.03	32	twitter	+	275.1	7	hi	+	1104.56	32	how	+	323.24
8	am	+	606.95	33	too	+	271.83	8	do	+	1007.11	33	got	+	320.55
9	for	+	605.85	34	site	+	262.84	9	just	+	952.91	34	sorry	+	319.46
10	ok	+	604.98	35	also	+	257.63	10	get	+	924.33	35	credit	+	314.13
11	my	+	558.17	36	online	+	257.1	11	my	+	806.99	36	hey	+	313.26
12	london	+	555.45	37	Reading	+	256.76	12	so	+	705.34	37	2011	+	313.21
13	to	+	539.45	38	news	+	252.37	13	me	+	697.28	38	ur	+	311.03
14	will	+	472.97	39	google	+	246.28	14	xx	+	680.8	39	thats	+	285.83
15	media	+	436.85	40	#socialmedia	+	246.28	15	ok	+	613.05	40	sure	+	276.94
16	lol	+	429.69	41	best	+	245.2	16	hope	+	608.04	41	twitter	+	272.84
17	hope	+	414.76	42	with	+	243.5	17	am	+	585.63	42	credit_card	+	262.16
18	social	+	406.44	43	credit_card	+	237.57	18	if	+	576.81	43	Reading	+	254.65
19	im	+	380.1	44	as	+	236.2	19	london	+	550.87	44	google	+	244.26
20	UK	+	369.43	45	xxx	+	233.18	20	please	+	541.57	45	#socialmedia	+	244.26
21	etc	+	361.18	46	blog	+	233.18	21	your	+	540.65	46	guys	+	239.93
22	post	+	349.12	47	posted	+	225.53	22	good	+	518.75	47	media	+	236.23
23	credit	+	348.41	48	check_out	+	220.45	23	really	+	479.31	48	etc	+	235.04
24	ur	+	340.61	49	learning	+	217.1	24	let_know	+	426.62	49	site	+	231.38
25	2011	+	335.37	50	new	+	212.93	25	lol	+	426.15	50	xxx		231.26

Figure 5: Key words and clusters used in the CANELC corpus, compared with spoken and written extracts from the BNC

Again, *I* is overused in CANELC in comparison to the written BNC data (with a log-likelihood of +5,219.64), but there is no significant difference in usage between the CANELC and the spoken BNC data. *You* (rated fourth, here, with a log-likelihood of +3,242.12) and other personal pronouns were all shown to be key words in comparison to the written element but their use was not as statistically different from the use in the spoken BNC data.

Terms related to the general thematic grouping 'information technology and computing' and 'the media, TV radio and cinema' (as characterised by the semantic tagging functionality in Wmatrix) such as *Google*, *site*, *Twitter*, *blog*, *social*, *media*, *BBC* and *socialmedia* are also shown to be more common in CANELC than its BNC counterparts. Similarly, references to communicating in digital environments are also particularly common in CANELC, with *fan*, *posted*, *news* and *learning* all featured in the top fifty key words. These latter terms can be broadly categorised under the thematic groupings of 'IT', 'the Media', 'telecommunication' and 'paper documents and writing' (which also includes terms such as *print*, *register*, *delete* and *list*) – themes which feature significantly more frequently in CANELC than the other corpora.

Figure 6 reveals some of the other key differences in the semantic categories (based on keywords and phrases used in the data) that are used at a significantly higher rate in the CANELC data, compared to the spoken

No.	Rel. freq. in CANELC	Rel. freq. in Written BNC	LL	difference	Category	Examples
1	10.51	7.44	+	2866.51	Pronouns	Could *you* please update me with....
2	1.37	0.45	+	2814.1	Discourse Bin	*Do you think* the academy logo looks ok?
3	0.21	0.01	+	1303.77	Polite	*Thanks* so much everyone for your hard work...
4	1.01	0.45	+	1250.03	Cause&Effect/Connection	Unfortunately, the *reason* why it isn't working...
5	0.7	0.3	+	931.91	Evaluation: Good	Thanks for the *super* prompt reply...
6	0.93	0.5	+	717.01	Time: Future	*Will* pay up this week my lovely ...
7	0.72	0.38	+	595.19	Paper documents and writing	you may have to get *tickets* in advance...
8	0.32	0.12	+	564.1	Information technology and computing	I will *upload* an example for you to look at...
9	0.47	0.21	+	558.46	If	I may hav an egg *if* u and dad can spare one?
10	0.24	0.07	+	527.95	Expected	I'm *hopefully* back at work tomorrow where....
11	0.35	0.14	+	514.15	Like	We *like* the mirror effect on its own.
12	0.17	0.04	+	507.75	Telecommunications	Please pick up your *phone* and vote on Sunday
13	0.52	0.28	+	415.92	Exclusivizers/particularizers	We have *just* been relaxing and catching up...
14	1.45	1.02	+	412.48	Location and direction	I'm having a quiet night *in* with the family.
15	0.56	0.31	+	409.4	Time: Present; simulateous	It's my nephews 4th birthday *today*. 20 mini-....

No.	Rel. freq. in CANELC	Rel. freq. in Spoken BNC	LL	difference	Category	Examples
1	1.01	0.19	+	3692.79	Cause&Effect/Connection	See above
2	0.92	0.36	+	1458.82	Geographical names	I am still heading to *Berlin* and really looking.....
3	0.21	0.01	+	1410.74	Polite	See above
4	0.32	0.05	+	1360.18	Information technology and computing	See above
5	0.24	0.03	+	1211.17	Expected	See above
6	0.72	0.35	+	762.57	Paper documents and writing	See above
7	0.2	0.04	+	645.99	Evaluation: Good	See above
8	0.27	0.08	+	643.44	Happy	U sound *happy*, yey :-) think. Think I'm gonna...
9	0.76	0.42	+	554.57	Objects generally	I worked on the *turnstiles* when they played at...
10	0.55	0.27	+	554.44	Other proper names	a sit down protest in *Vodaphone* the other day...
11	0.32	0.13	+	491.13	Money and pay	I probably can't really *afford* it until then anyway
12	0.1	0.01	+	464.15	The Media	You should *publish* a book with that tweet
13	0.25	0.09	+	457.21	Investigate, examine, test, search	Just *checking*. Need the cash that's all...
14	1.11	0.73	+	433.55	Time: Period	has been going on for some *years* and there is....
15	0.93	0.59	+	430.8	Speech acts	list but I also *recommend* that you attend every....

Figure 6: Comparing semantic categories of the CANELC data versus the spoken and written BNC data

and written elements of the BNC (based on the UCREL[14] semantic analysis system, as featured in Wmatrix; see Wilson and Rayson, 1993).

Content related to time and place (including the categories: 'geographical names', 'time: period', 'time: future', 'time: present/simultaneous' and 'location and direction') also feature more frequently in the CANELC corpus compared to the spoken and written data. In Figure 4 we saw that clusters such as 'last night', 'next week', 'next year', 'this morning' and 'at the moment', in particular, are especially ubiquitous. This is an interesting finding because although e-language is actually asynchronous, there may only be a short delay between the time that e-language is composed and the time that it is read and responded to.

The use of these temporal deictic markers (as with the use of personal pronouns), suggests forms of communication that allow for an immediate or near-immediate information exchange, a forum for communicating reports of events and incidents in near real time, as the understanding of the temporal referent is shared. In fact, on closer inspection of some of the Twitter, e-mail and SMS data in particular, messages were often sent by users and then responded to within hours, even minutes, closing the gap between

[14] The University Centre for Computer Corpus Research on Language, Lancaster University.

Word/cluster	Frequency
thanks	669
grateful	21
thank	7
thank u	14
thankyou	9
compliment	7
courtesy	6
refined	5
thank-you	3
complimentary	3
polite	2
good_manners	1
apologetic	1
politely	1
	749

Figure 7: Common politeness terms use in CANELC

production and reception, and possibly accounting for the differences observed. However, in contrast with face-to-face, spoken discourse, the actual physical space is rarely shared at the point when the message is sent. The lack of shared physical space may lead to an over-compensation in the use of deictic markers, as a means of establishing and reconfirming a shared 'digital space' between senders and recipients. Such reconfirmation is not a required part of spoken interaction as the social, physical and temporal context is frequently changeable.

Aside from deictic markers, Figure 6 also reveals that the use of politeness strategies is also more frequent in CANELC than the BNC data, with log-likelihood score of +1,410.74 (a frequency of 103) compared to the spoken BNC and +1,303.77 when compared to the written BNC (a frequency of 130). As seen in Figure 7, *thank* appears to be a particularly common word in e-language, occurring 669 times across the corpus.

This frequent use of politeness terms is seen in all sub-types of the CANELC data, with the language of the e-mails ranking as having a particularly high number of politeness terms and the blogs with the least (although even for blogs, the number is still significantly higher than what is seen in the BNC). This finding mirrors that seen by Herring who found that 'public CMD tends to be less polite than private CMD' (2002: 19), although this obviously depends on the purpose of communication, who the message is intended for (i.e., whether it is aimed at a specific person or a group of people) and the nature of the relationship between the sender and sendee. The fact that blogs, Twitter exchanges and much of the discussion board content included in CANELC is publicly accessible suggests that the maintenance of face and positive politeness are critical ingredients for maximising the number of people that will follow one's online existence. This would help to explain the frequent use of politeness strategies across all the modes of e-language

examined here. However, specific conventions for doing this 'successfully' need to be examined in more detail.

Another interesting feature of e-Language, which is used more extensively than in spoken and written data, is a closing with kisses: *x*, *xx* and *xxx*. The average length is a single *x*, unless the recipient is defined as a close friend or partner, and there was also evidence of the use of *x* between colleagues and friends of the same and different genders (for both men and women), a device which thus seems to be accepted conventionally for use by all. The use of *x* is seen to be highly frequent in all of the modes of e-language; most commonly featuring in SMS messages and least frequently in blog data. *X* broadly functions as a relationship maintenance device – a method of bringing the sender and recipient of a message closer together, again despite the physical distance. It acts almost as a signing off method that is more personal and expressive than a full-stop or a signature. Again, a more detailed exploration of the differences in usage of *x* across the different e-language modes and individual users is something that will be explored further in the future. Questioning what precedes or follows a message with an *x*, and questioning the function of a message will also help us to construct a more detailed understanding of this feature. For example, compare the following two SMSs from CANELC:

(SMS.224)
How did the footie go? U watching that drama on 4? Very sad :-(. . . .
Filmed in notts x

(SMS.3964)
Its just a copy of wots there, theyre usingthe old bits as a template. All in 8×3. They know all this.

The function of the SMS.3964, sent by a manager to a colleague, is purely transactional (a form of information provision), while the second example is of a more intimate kind (an information request but in a more socialising capacity). For SMS.224, the *x* acts to maintain and reinforce the relationship between sender and recipient. This is less critical in the second message.

9. Summary

This paper has introduced the one-million word CANELC corpus. It details how the corpus was constructed and illustrates how it may be used to help us to examine the structure and use of language in digital environments with, as can be seen from the corpus construction, opportunities to examine how e-language varies across different domains, across different levels of formality and with particular attention to the spokenness and high levels of interactivity of some e-communication. While further research into discourse

within digital domains needs to be carried out, we believe that CANELC provides us with some of the main foundations for doing this.

This opens the door to a variety of interesting questions about the use of language in digital contexts, questions that, with time, we hope to explore using CANELC. Among the possibilities are, at a more micro level, analysis of further seemingly e-specific forms such as politeness phenomena and deixis across the database, as well as exploration of key recorded forms of spoken grammar which are outlined by major grammars such as Biber *et al*. (1999) and Carter and McCarthy (2006) and include vagueness markers, ellipsis, modal expressions, fronting, headers and tails At a more macro level, the possibilities include: fuller comparisons between CANELC data and other e-language corpora; the collection and analysis of Facebook data to explore the nature and the extent of linguistic differences and distinctions between this popular medium and other e-language forms; extending CANELC to embrace a larger range of e-mail data from a wider variety of contexts of use; examining the extent to which spokenness – not just in e-communication but in writing in general – is before our very eyes both a growing phenomenon and a significant part of systemic contemporary language change.

References

Allan, S. 2006. Online News: Journalism and the Internet. Maidenhead, UK: Open University Press.

Atkins, S. 2011. A Cognitive Linguistic Perspective on Social Space in Online Health Communities. Unpublished PhD Thesis. Nottingham: University of Nottingham.

Baron, N. 1998. 'Writing in the age of email: the impact of ideology versus technology', Visible Language 32 (1), pp. 35–53.

Baron, N. 2000. Alphabet to Email: How Written English Evolved and Where it's Heading. London: Routledge.

Baron, A. and P. Rayson. 2008. 'VARD 2: a tool for dealing with spelling variation in historical corpora' in proceedings of the Postgraduate Conference in Corpus Linguistics. 22 May 2008. Aston University, Birmingham, UK.

Bates, E. 1976. Language and Context. New York: Academic Press.

Baym, N. 1995. 'The emergence of community in computer-mediated communication' in S.G. Jones (ed.) Cybersociety: Computer-mediated Communication and Community, pp. 138–63. Thousand Oaks, California: Sage.

Beißwenger, M. 2007. Sprachhandlungskoordination in der Chat-Kommunikation (Linguistik – Impulse and Tendenzen 26). Berlin and New York: de Gruyter.

Biber, D. 1992. 'On the complexity of discourse complexity: a multi-dimensional analysis', Discourse Processes 15 (2), pp. 133–63.

Biber, D. 1993. 'Representativeness in corpus design', Literary and Linguistic Computing 8 (4), pp. 243–57.

Biber, D., S. Conrad, G. Leech, J. Svartvik and E. Finegan. 1999. The Longman Grammar of Spoken and Written English. Harlow: Longman.

Borau, K., C. Ullrich, J. Feng and R. Shen. 2009. 'Microblogging for language learning: using Twitter to train communicative and cultural competence', ICWL 2009, LNCS 5686, pp. 78–87.

Boyd, D. and J. Heer. 2006. 'Profiles as conversation: networked identity performance on Friendster' in proceedings of the Hawai'i International Conference on System Sciences (HICSS-39), Persistent Conversation Track. 4–7 January 2006. Kauai, Hawai'i: IEEE Computer Society.

Brown, G. 1989. 'Making sense: the interaction of linguistic expression and contextual information', Applied Linguistics 10 (1), pp. 97–108.

Carter, R.A. and M.J. McCarthy. 2006. Cambridge Grammar of English. Cambridge: Cambridge University Press.

Chafe, W.L. and J. Danielewicz. 1987. 'Properties of spoken and written language' in R. Horowitz and S.J. Samuels (eds) Comprehending Oral and Written Language, pp. 83–113. New York: Academic Press.

Collot, M. and N. Belmore. 1996. 'Electronic language: a new variety of English' in S.C. Herring (ed.) Computer-Mediated Communication: Linguistic, Social and Cross-cultural Perspectives, pp. 13–28. Amsterdam: John Benjamins.

Condon, S.L. and C.G. Cech. 2001. 'Profiling turns in interaction' in proceedings of the thirty-fourth Annual Conference of the Hawai'i International Conference on System Sciences. Los Alamitos, California: IEEE Computer Society Press.

Crystal, D. 1998. Language Play. Cambridge: Cambridge University Press.

Crystal, D. 2001. Language and the Internet. Cambridge: Cambridge University Press.

Crystal, D. 2003. 'The joy of text', Spotlight Magazine, pp. 16–17.

Crystal, D. 2008. Txtng: The Gr8 Db8. Oxford: Oxford University Press.

Danet, B. 2002. 'The language of email'. Lecture II. European Union Summer School lecture. June 2002. Rome: University of Rome.

Döring, N. 2002. '1 Brot, Wurst, 5 Sack Äpfel I.L.D – Kommunikative Funktionen von Kurzmitteilungen (SMS)' [1 bread, sausage, 5 bags of apples I.L.Y.' – Communicative functions of text messages (SMS)], Zeitschrift für Medienpsychologie 14 (3), pp. 118–28.

Duranti, A. and C. Goodwin (eds). 1992. Rethinking Context: Language as an Interactive Phenomenon. Cambridge: Cambridge University Press.

Faulkner, X. and F. Culwin. 2005. 'When fingers do the talking: a study of text messaging', Interacting with Computers 17 (2), pp. 167–85 .

Forsyth, E.N. and C.H. Martell. 2007. 'Lexical and discourse analysis of online chat dialog' in proceedings of the first IEEE International Conference on Semantic Computing (ICSC 2007). 19–26 September 2007. Irvine, California.

Gillmor, D. 2004. We the Media: Grassroots Journalism By the People, For the People. Farnham, Surrey: O'Reilly Media.

Green, L.J. 2002. African American English: A Linguistic Introduction. Cambridge: Cambridge University Press.

Grinter, R.E. and M. Eldridge. 2003. "Wan2talk? Everyday text messaging' in proceedings of the ACM Conference on Human Factors in Computing Systems (CHI), pp. 441–8 .

Halliday, M.A.K. and R. Hasan. 1989. Language, Context and Text: Aspects of Language in a Social Semiotic Perspective. Oxford: Oxford University Press.

Hård af Segerstad, Y. 2002. Use and Adaptation of the Written Language to the Conditions of Computer-Mediated Communication. Unpublished PhD thesis. Goteborg: University of Goteborg.

Herring, S.C. 2002. 'Computer-mediated communication on the Internet', Annual Review of Information Science and Technology 36 (1), pp. 109–68 .

Herring, S.C. 2007. 'A faceted classification scheme for computer-mediated discourse', Language@Internet 4 (1), pp. 1–37 .

Heylighen, F. and J.-M Dewaele. 2003. 'Variation in the contextuality of language: an empirical measure', Foundations of Science 7 (3), pp. 293–340 .

Honeycutt, C. and S.C. Herring. 2009. 'Beyond microblogging: conversation and collaboration via Twitter' in proceedings of the Forty-Second Hawai'i International Conference on System Sciences (HICSS-42). Los Alamitos, California: IEEE Press.

Horn, C., O. Pimas, M. Granitzer, E. Lex and K.-C. Graz. 2011. 'Realtime ad hoc search in Twitter: know-center at TREC Microblog Track 2011' in proceedings of the twentieth Text ReEtrieval Conference (TREC 2011). 15–18 November 2011. Gaithersburg, Maryland.

Hymes, D. 1974. Foundations in Sociolinguistics: An Ethnographic Approach. Philadelphia: University of Pennsylvania Press.

Jansen, B.J., M. Zhang, K. Sobel and A. Chowdury. 2009. 'Twitter power: tweets as electronic word of mouth', Journal of the American Society for Information Science and Technology 60 (11), pp. 2169–88.

Jones, Q. 1997. 'Virtual-communities, virtual settlements and cyber archaeology: a theoretical outline', Journal of Computer Mediated Communication 329 (3), online.

Klimt, B. and Y. Yang. 2004. 'Introducing the Enron Corpus' in proceedings of CEAS 2004 – First Conference on Email and Anti-Spam, pp. 30–31. Mountain View, California, USA.

Ko, K. 1996. 'Structural characteristics of computer-mediated language: a comparative analysis of InterChange discourse', Electronic Journal of Communication 6 (3). Available online, at: http://www.cios.org/www/ejc/v6n396.htm

Labov, W. 1972. Sociolinguistic Patterns. Philadelphia, Pennsylvania: University of Pennsylvania Press.

Leech, G. 2000. 'Grammar of spoken English: new outcomes of corpus-oriented research', Language learning 50 (4), pp. 675–724.

Ling, R. 2003. 'The socio-linguistic of SMS: an analysis of SMS use by random sample of Norwegians' in R. Ling and P. Pedersen (eds) Mobile Communications: Renegotiation of the Social Sphere, pp. 335–49. London: Springer.

Myers, B. 2010. 'Theology 2.0: blogging as theological discourse', Cultural Encounters 6 (1), pp. 47–60.

Myers, G. 2010. The Discourse of Blogs and Wikis. London: Continuum.

Nelson, K., S. Engel and A. Kyratzis. 1985. 'The evolution of meaning in context', Journal of Pragmatics 9 (4), pp. 453–74.

Oksman, V. and J. Turtianen. 2004. 'Mobile communication as a social stage: meanings of mobile communication in everyday life among teenagers in Finland', New Media and Society 6 (3), pp. 319–39.

Orasan, C. and R. Krishnamurthy. 2002. 'A corpus-based investigation of junk emails' in proceedings of LREC-2002. Las Palmas, Spain.

Panteli, N. 2002. 'Richness, power cues and email text', Information and Management 40 (2), pp. 75–86.

Puschmann, C. 2009. 'Diary or Megaphone? The pragmatic mode of weblogs', paper presented at Language in the (New) Media: Technologies and Ideologies. 3–6 September 2009. Seattle, Washington, USA.

Rayson, P. 2003. Matrix: A Statistical Method and Software Tool for Linguistic Analysis through Corpus Comparison. Unpublished PhD thesis. Lancaster: Lancaster University.

Rheingold, H. 1993. The Virtual Community: Homesteading on the Electronic Frontier. New York: HarperCollins.

Schler, J., M. Koppel, S. Argamon and J. Pennebaker. 2006. 'Effects of age and gender on blogging' in proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs.

Scollon, R. and S. Scollon. 2003. Discourse in Place: Language in the Material World. London: Routledge.

Shortis, T. 2007. 'Gr8 txtpectations: the creativity of text spelling', English Drama Media Journal 8, pp. 21–6.

Sutherland, J. 2002. 'Cn u txt?'. Featured in *The Guardian*, 11 November.

Tagg, C. 2009. A Corpus Linguistics Study of SMS Text Messaging. Unpublished PhD Thesis. Birmingham: University of Birmingham.

Thurlow, C. and A. Brown. 2003. 'Generation txt? Exposing the sociolinguistics of young people's text-messaging', Discourse Analysis Online 1 (1). Available online, at: http://extra.shu.ac.uk/daol/

Widdowson, H.G. 1998. 'Communication and community: the pragmatics of ESP', English for Specific Purposes 17 (1), pp. 3–14.

Wilson, A. and P. Rayson. 1993. 'Automatic content analysis of spoken discourse' in C. Souter and E. Atwell (eds) Corpus-based Computational Linguistics, pp. 215–26. Amsterdam: Rodopi.

Zappavingna, M. 2011. Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web. London: Continuum.